

# Dynamics of Gene Duplication in the Genomes of Chlorophyll *d*-Producing Cyanobacteria: Implications for the Ecological Niche

Scott R. Miller<sup>\*,1</sup>, A. Michelle Wood<sup>2,3</sup>, Robert E. Blankenship<sup>4,5</sup>, Maria Kim<sup>6</sup>, and Steven Ferriera<sup>†,6</sup>

<sup>1</sup>Division of Biological Sciences, The University of Montana

<sup>2</sup>Department of Biology, University of Oregon

<sup>3</sup>NOAA Atlantic Oceanographic and Meteorological Laboratory, Miami, Florida

<sup>4</sup>Department of Biology, Washington University, St. Louis, Missouri

<sup>5</sup>Department of Chemistry, Washington University, St. Louis, Missouri

<sup>6</sup>J. Craig Venter Institute, Rockville, Maryland

<sup>†</sup>Present address: Broad Institute, Cambridge, Massachusetts

\*Corresponding author: E-mail: scott.miller@umontana.edu.

**Accepted:** 7 June 2011

**Data deposition:** This Whole Genome Shotgun project has been deposited at DNA Data Bank of Japan/EMBL/GenBank under the accession AFEJ00000000. The version described in this paper is the first version, AFEJ01000000

## Abstract

Gene duplication may be an important mechanism for the evolution of new functions and for the adaptive modulation of gene expression via dosage effects. Here, we analyzed the fate of gene duplicates for two strains of a novel group of cyanobacteria (genus *Acaryochloris*) that produces the far-red light absorbing chlorophyll *d* as its main photosynthetic pigment. The genomes of both strains contain an unusually high number of gene duplicates for bacteria. As has been observed for eukaryotic genomes, we find that the demography of gene duplicates can be well modeled by a birth–death process. Most duplicated *Acaryochloris* genes are of comparatively recent origin, are strain-specific, and tend to be located on different genetic elements. Analyses of selection on duplicates of different divergence classes suggest that a minority of paralogs exhibit near neutral evolutionary dynamics immediately following duplication but that most duplicate pairs (including those which have been retained for long periods) are under strong purifying selection against amino acid change. The likelihood of duplicate retention varied among gene functional classes, and the pronounced differences between strains in the pool of retained recent duplicates likely reflects differences in the nutrient status and other characteristics of their respective environments. We conclude that most duplicates are quickly purged from *Acaryochloris* genomes and that those which are retained likely make important contributions to organism ecology by conferring fitness benefits via gene dosage effects. The mechanism of enhanced duplication may involve homologous recombination between genetic elements mediated by paralogous copies of *recA*.

**Key words:** *Acaryochloris*, *recA*, homologous recombination, plasmid.

## Introduction

Gene duplication is an important mechanism of gene innovation and genome evolution (Ohno 1970; Taylor and Raes 2004). A substantial fraction of eukaryotic, bacterial, and archaeal genomes may be composed of divergent paralogs resulting from gene family expansion (Coissac et al. 1997; Jordan et al. 2001; Gevers et al. 2004; Makarova et al.

2005), and examples of the role of gene duplicates as a source of raw material for the origin of evolutionary novelties and diversification abound (e.g., True and Carroll 2002; Irish and Litt 2005; Wagner 2008).

In addition to ancient paralogs, eukaryotic genomes generally contain a large number of recent duplicates (Lynch and Conery 2000, 2003). By contrast, although gene duplications can occur at frequencies as high as  $10^{-3}$  per gene per

generation in bacterial genomes (Anderson and Roth 1977; Haack and Roth 1995; Reams et al. 2010), these duplicates are quickly purged from the genome unless they confer fitness advantages via dosage effects (i.e., enhanced gene expression; Roth et al. 1996; Romero and Palacios 1997; Reams et al. 2010). Consequently, bacterial genomes typically harbor few recent duplicates (Hooper and Berg 2003b).

Here, we analyzed the age distributions and selection histories of duplicate genes in the genomes of two strains of the cyanobacterium *Acaryochloris* which contain an unusually large number of recent (i.e., low divergence) duplicates for bacterial genomes: the previously finished genome of *Acaryochloris* strain MBIC11017 (Swingley et al. 2008) and a draft genome that we have assembled for *Acaryochloris* strain CCME 5410. *Acaryochloris* spp. specialize on far-red wavelengths of solar radiation that are inaccessible to other photosynthetic organisms through their unique ability to produce chlorophyll (Chl) *d*, a structural relative of Chl *a*, as the major pigment in photosynthesis (Miyashita et al. 1996; Miller et al. 2005). This recently discovered group has been detected in diverse marine, freshwater, and terrestrial habitats (Behrendt et al. 2011) and may make a significant contribution to the global carbon cycle (Kashiyama et al. 2008). Strain MBIC11017 was isolated from the Great Barrier Reef (Miyashita et al. 1996), where *Acaryochloris* biofilms commonly develop underneath ascidians (Kühl et al. 2005). Strain CCME 5410 was isolated from a benthic epilithic biofilm in the Salton Sea (Miller et al. 2005), a saline, eutrophic closed basin lake in southern California with major inputs from agricultural runoff and municipal wastewater.

We report that rates of duplication and duplicate loss fall within the range of values estimated for eukaryotic rather than bacterial genomes. Although duplicates may experience a brief period of relaxed selection, most are rapidly lost from the genome, and those which are retained are subject to strong purifying selection. The idiosyncratic duplicate pools of the respective genomes include many open reading frames (ORFs) that appear to be important for fitness in the specific environments from which the strains were derived, including a large number of duplicates involved in iron acquisition in strain MBIC11017 and an enrichment of duplicated loci involved in heavy metal resistance in strain CCME 5410. We conclude with consideration of the mechanisms which may contribute to the unusual duplication dynamics of these bacteria.

## Materials and Methods

### *Acaryochloris* Strain CCME 5410 Genome

Cells were grown, and genomic DNA was isolated as previously described (Swingley et al. 2008). The CCME 5410 genome was sequenced on the 454 FLX Titanium platform and assembled with Roche's Newbler de novo assembler with default overlap settings. The JCVI auto-annotation pipeline

was used to identify sequence features and assign functional annotation. Protein-coding sequences were predicted with Glimmer3 (Delcher et al. 1999), tRNAs were identified with the tRNAscan tool (Lowe and Eddy 1997), and rRNA genes and other structural RNAs were identified directly from Blast (Altschul et al. 1990) matches to Rfam. Functional annotation of proteins was assigned based on coding sequences comparison against the CHAR database of experimentally verified proteins and functional annotations, TIGRFAM (Haft et al. 2003) and Pfam (Finn et al. 2008) protein family databases, the PANDA repository of nonredundant protein and nucleotide data, and by computationally derived assertions including lipoprotein and transmembrane helix signatures. Assembled contigs greater than 5 kbp in length were assigned to chromosome or plasmid elements by a nucmer alignment against the *Acaryochloris marina* strain MBIC11017 reference genome in the MUMmer package (Kurtz et al. 2004).

This whole-genome shotgun project has been deposited at DNA Data Bank of Japan/EMBL/GenBank under the accession AFEJ000000000. The version described in this paper is the first version, AFEJ010000000.

### Identification and Analysis of Recent Duplicates

Paralogs within the genomes of *Acaryochloris* strains CCME 5410 and MBIC11017 were identified by local BlastP searches (Altschul et al. 1990) of each inferred protein sequence against its genome. Because the study was focused on the pools of recent duplicates, putative paralogs sharing less than 50% amino acid identity were removed from the data set. A similar search strategy was used to identify shared duplicates via reciprocal local BlastP searches. ORFs annotated as transposases, integrases, or identified as having significant homology ( $E < 0.05$ ) to insertion sequence (IS) elements by BlastP against the IS Finder database ([www-is.biotoul.fr/is.html](http://www-is.biotoul.fr/is.html)) were also removed, as were gene families with more than ten paralogs (typically transposases).

Nucleotide alignments of duplicates were obtained by the manual adjustment of ClustalW automated alignments (Thompson et al. 1994) using the amino acid alignments as a guide. Silent site divergence ( $d_s$ ) and replacement site divergence ( $d_n$ ) between aligned nucleotide sequences of duplicate pairs were estimated by the maximum likelihood (ML) procedure implemented in the CODEML program of the PAML software package (version 3.14; Yang 1997). For all models, codon usage (the average nucleotide frequencies at the three codon positions) and transition/transversion bias were estimated from the data. Only duplicate pairs with  $d_s < 5$  were considered for further analysis.

Most cases involved a duplicate pair resulting from a single duplication event. For cases involving more than two paralogs, we used phylogenetics to distinguish the duplication events (e.g., resolution of three duplicates by reconstruction of the two duplication events). Phylogenies of aligned

nucleotide sequences were inferred by ML with PAUP\* (Swofford 1996) according to a model of DNA sequence evolution selected by hierarchical likelihood ratio tests implemented by Modeltest (Posada and Crandall 1998). For the ML heuristic search, a starting tree was obtained by random sequence addition, and branch swapping was performed by tree bisection and reconnection. The resulting topology was used to specify the tree for the PAML model as described above.

### **recA Phylogeny Reconstruction and Tests of Protein Adaptation**

Nucleotides (1,023) of the *recA* genes of *Acaryochloris* and other representative cyanobacteria were aligned by ClustalW (Thompson et al. 1994). A ML tree was reconstructed with PAUP\* as described above according to the general time reversible (GTR) + I + G model of sequence evolution selected by Modeltest (Posada and Crandall 1998) and bootstrapped 100 times. A Bayesian analysis was performed with MrBayes (Huelsenbeck and Ronquist 2001) using the GTR + I + G model. Two independent chains of 1,000,000 generations of Markov chain Monte Carlo were analyzed, with trees sampled every 1,000 generations. Chain convergence was evaluated by the average standard deviation of split frequencies, and the first 20% of trees were discarded as burn-in. To test for the signature of positive selection during *Acaryochloris recA* diversification, branch-site models of codon evolution (Yang and Nielsen 2002) were implemented with PAML. Likelihood scores of nested models which either allow for a class of positively selected codon sites (i.e.,  $d_N/d_S > 1$ ) or constrain  $d_N/d_S$  to be less than or equal to 1 (the nearly-neutral model) were compared with a  $\chi^2$  test. For branches of the *recA* tree for which the nearly-neutral model was rejected, a Bayes empirical Bayesian analysis was used to infer which codon sites belonged to the positively selected class with high (>95%) posterior probability.

## **Results and Discussion**

### **Acaryochloris Strain CCME 5410 Genome**

The *Acaryochloris* strain CCME 5410 genome was pyrosequenced to approximately 24× coverage depth, and the resulting genome data assembled into 511 contigs greater than 500 bp, with an  $N_{50}$  of 37,625 bp. The estimated genome size of 7.88 Mbp is somewhat smaller than that of the previously finished genome of *Acaryochloris* strain MBIC11017 (table 1; Swingley et al. 2008) as well as of a recently described strain isolated from the Great Barrier Reef for which an unpublished draft genome sequence has been obtained (~8.37 Mbp; Mohr et al. 2010) but is still considerably larger than those of other unicellular cyanobacteria. The strain CCME 5410 genome likewise contains fewer predicted ORFs than that of strain MBIC11017. The two ge-

**Table 1**

General Features of the CCME 5410 and MBIC11017 Genomes

	CCME 5410	MBIC11017 <sup>a</sup>
Genome size (Mbp)	7.88	8.36
GC content (%)	47	47
ORFs	8383	8528
Strain-specific ORFs	2261	2406
IS elements	552	487

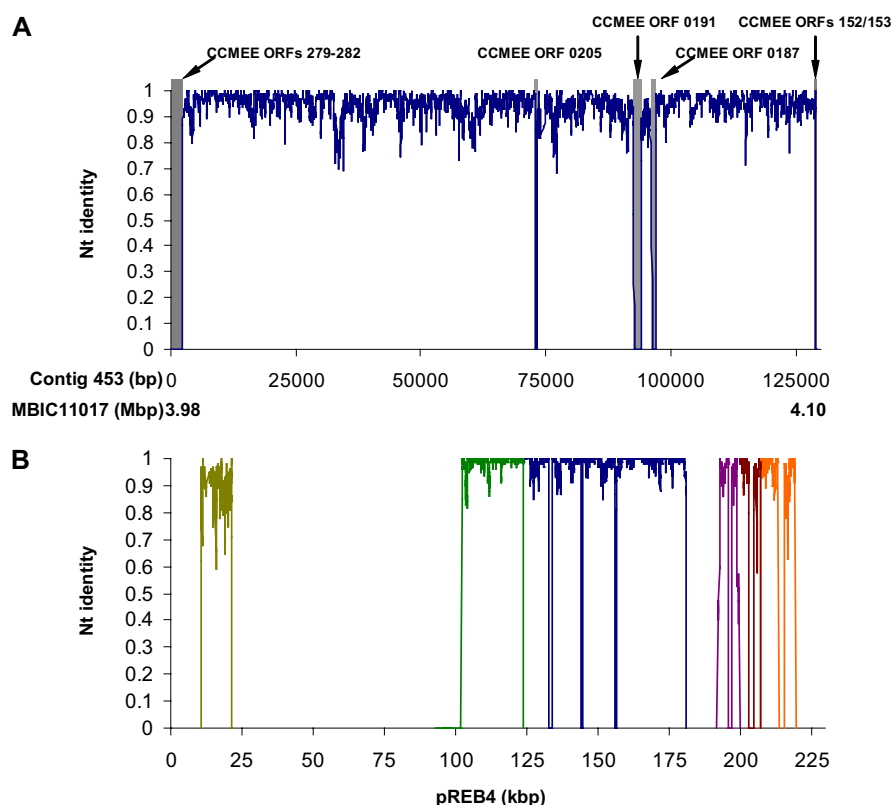
<sup>a</sup> Data from Swingley et al. (2008).

nomes share similar base composition and a high number of ORFs with significant homology to IS elements (table 1).

The CCME 5410 and MBIC11017 genomes share 6,122 putative orthologs, with greater than 25% of predicted ORFs in each genome absent from the other (table 1). For the closed MBIC11017 genome, we can identify with certainty the genetic element on which each of these idiosyncratic ORFs resides. In addition to a circular chromosome, it contains nine apparently single-copy plasmids, varying in size from approximately 2.1 to 374 kbp, which together comprise roughly 22% of the genome (Swingley et al. 2008). For the CCME 5410 assembly, we provisionally assigned contigs greater than 5 kbp in length to either the chromosome or a plasmid element using a nucmer alignment against the MBIC 11017 genome (supplementary table S1, Supplementary Material online). This length cutoff was chosen because most short contigs either exhibited no homology to the MBIC11017 genome and/or encoded an IS element(s). One hundred and eighty-eight contigs totaling 5.81 Mbp were assigned to the chromosome, and 61 contigs with a cumulative size of 1.52 Mbp were assigned to plasmids (supplementary table S1, Supplementary Material online).

Gene content is generally conserved on the two *Acaryochloris* chromosomes. Approximately, 89% of ORFs on the MBIC11017 chromosome (5,621/6,342) have homologs in the CCME 5410 genome, whereas 83.5% of ORFs (4,951/5,932) on contigs assigned to the CCME 5410 chromosome have homologs in the MBIC11017 genome. Mapping of these chromosome contigs to the MBIC11017 reference indicated a high degree of sequence conservation and local synteny between chromosomes (fig. 1A; reference range data in supplementary table S1, Supplementary Material online).

By contrast, differences in gene content between the genomes are concentrated on plasmids. Seventy-seven percent (1,685/2,186) of MBIC11017 plasmid ORFs have no homolog in the CCME 5410 genome, accounting for 70% of the ORFs absent from the latter. The individual plasmids vary in the fraction of ORFs with homologs in the CCME 5410 genome from 0% (pREB9) to ~48% (pREB4). Similarly, for CCME 5410 contigs assigned to a plasmid, 55% of the 1,649 ORFs lacked a homolog in the MBIC11017 genome. In addition, few large blocks of synteny were observed among the MBIC11017 plasmid ORFs



**FIG. 1.**—(A) Sequence and gene order conservation between *Acaryochloris* chromosomes illustrated for CCMEE 5410 contig 453, which maps to positions 3.98–4.10 Mbp on the strain MBIC11017 chromosome. The y axis is the probability that a pair of aligned nucleotides is identical in state between genomes along a sliding window of 100 nucleotide sites with a 25 site step-size. Nonhomologous regions include transposases at the contig breakpoints, CCMEE 5410 ORFs missing in the MBIC11017 genome (ORFs 191 and 205) and an ORF (ORF 187) which maps to coordinates 6.368–6.369 Mbp on the MBIC11017 chromosome. (B) Sequence conservation between *Acaryochloris* genomes for CCMEE 5410 contigs homologous to MBIC11017 plasmid pREB4: contig 468 (blue), contig 500 (green), contig 510 (orange), contig 511 (gold), contig 576 (brown), contig 598 (plum). Approximately, half of pREB4 is missing from the CCMEE 5410 genome. The fraction of each contig that maps to pREB4 ranges from 34.5% (contig 511) to 96% (contig 468).

that were shared between the genomes (supplementary table S1, Supplementary Material online). The most extensive syntenic regions were clustered on plasmid pREB4 in a region spanning MBIC 11017 ORFs D0134–D0214 (fig. 1B). Blocks of synteny from this plasmid include genes responsible for the biosynthesis and maturation of a bidirectional hydrogenase (ORFs D0176–D0197; nucleotides 140334–159433) and a complete set of loci encoding an alternative ATP synthase (ORFs D0157–D0167; nucleotides 123957–132033). These results suggest a greater instability of the *Acaryochloris* plasmids compared with the chromosome.

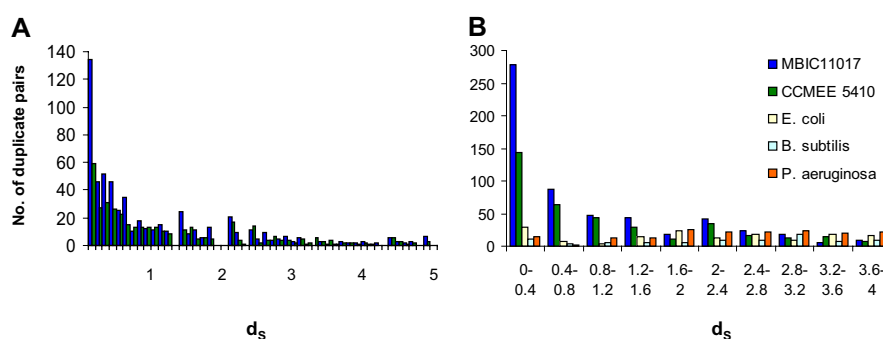
### Age Distribution of Duplicated Genes in *Acaryochloris* Genomes

Both genomes are notable for their large number of recent paralogs. We identified 393 and 597 duplicate pairs with synonymous-site divergence ( $d_s$ ) less than  $d_s = 5$  in the genomes of *Acaryochloris* strains CCMEE 5410 and MBIC11017, respectively. A majority of duplicated regions

involve only a single protein-coding ORF; only ~29% of pairs ( $N = 174$ ) in the strain MBIC11017 genome and ~35% of pairs ( $N = 136$ ) in the strain CCMEE 5410 genome were a part of duplicated blocks of greater than one ORF.

Most duplicates belong to the least divergent classes ( $d_s < 1$ ; fig. 2A). The difference between strains in the observed number of duplicate pairs is primarily due to a greater number of duplicates in these classes in the genome of strain MBIC11017, which contains approximately double the number of duplicate pairs with  $d_s < 0.5$  (278 vs. 143). By contrast, the number of duplicate pairs with  $d_s > 2$  is similar between the genomes (121 vs. 102). For both *Acaryochloris* genomes, the number of duplicate pairs with  $d_s < 1.5$  is very large compared with other representative bacterial genomes (fig. 2B; Hooper and Berg 2003b). For greater levels of  $d_s$ , duplicate numbers are comparable, with the exception of an apparent enhanced density of duplicates in *Acaryochloris* genomes centered on  $d_s$  values of ~2–2.4 (fig. 2).

Most duplicate pairs from the least divergent classes are strain specific, whereas more divergent duplicates are



**FIG. 2.**—(A) Frequency distributions of duplicate pairs for *Acaryochloris* strains MBIC11017 (blue) and CCME 5410 (green). (B) Frequency distributions of *Acaryochloris* duplicate pairs compared with data for *Escherichia coli* K12, *Bacillus subtilis* 168, and *Pseudomonas aeruginosa* PA01 (from Hooper and Berg [2003b]).

generally more likely to be present in both genomes (fig. 3). This pattern is in accord with the expectation that silent site divergence is generally a reasonable proxy for the age of a duplication event and that less divergent duplicate pairs have therefore largely originated following the divergence of these strain lineages from their common ancestor. However, there are a number ( $N = 60$ ) of low divergence ( $d_s < 1$ ) duplicate pairs in the strain CCME 5410 genome that also are found in the strain MBIC11017 genome. Such pairs may be the result of convergent duplication events following strain divergence or, alternatively, may appear “younger” than they are due to either gene conversion or extreme sequence conservation at synonymous sites. We believe that slower than average evolutionary rates is of primary importance for these loci because clear evidence from phylogenetic analyses for either convergent evolution or gene conversion (i.e., paralogs clustering by strain) was observed for only a minority ( $N = 14$ ) of duplicate pairs (data not shown). Among more divergent duplicates, approximately 50% (77/166) of duplicate pairs of divergence level  $d_s > 1$  in the strain CCME 5410 genome are present in the MBIC11017 genome (fig. 2). The unique duplicates among the more divergent age classes suggest that there has been

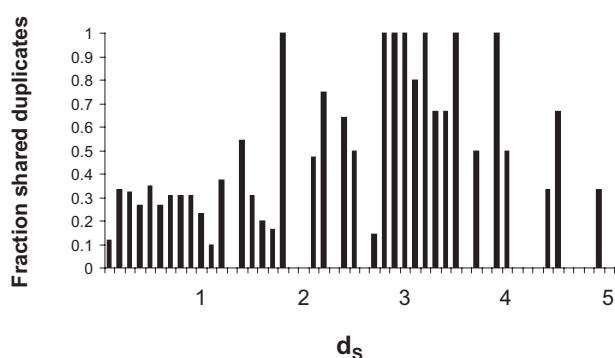
the differential retention of older duplicates between genomes following strain divergence.

### Estimation of Duplicate Birth and Death Rates

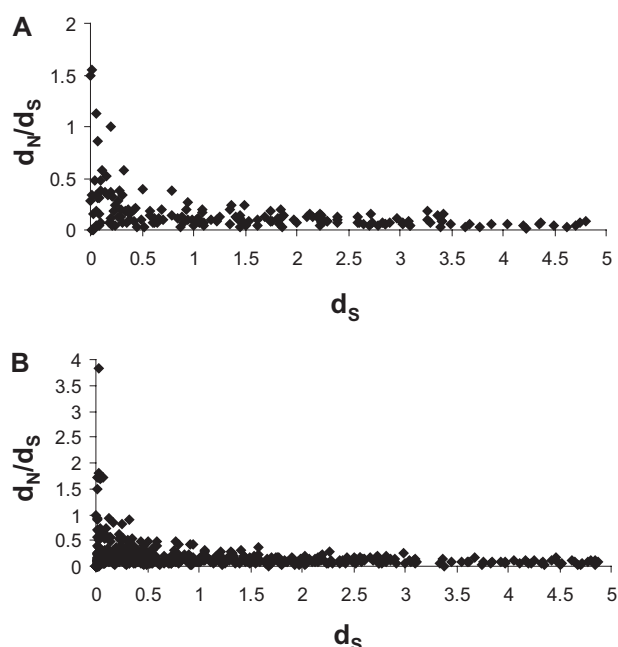
Following the approach of Lynch and Conery (2000, 2003), we modeled each age distribution as a steady-state birth–death process in order to estimate the rates at which duplicates arise and disappear from the respective *Acaryochloris* genomes. Because the assumption of constant birth and death rates is more likely to be valid over a short time scale, we limited the analyses to duplicate pairs with silent site divergence less than  $d_s = 0.1$ . For both data sets, we also excluded duplicate pairs in these age classes ( $N = 7$  pairs) found in both genomes (see above) to remove the potential impacts on the analysis of either gene conversion events or slowly evolving duplicates. We note that similar results were obtained for the full data set (not shown).

Under a steady-state birth–death process, the instantaneous rate of removal of duplicates from the genome ( $d$ ) can be estimated by the slope of the linear regression of  $\ln n_i$  on  $d_s$ , where  $n_i$  is the number of duplicate pairs in age class  $i$ . The regression models explained most of the variation in both data sets ( $R^2 = 0.82$ ,  $P < 0.0001$  for *Acaryochloris* strain CCME 5410;  $R^2 = 0.76$ ,  $P < 0.0001$  for *Acaryochloris* strain MBIC11017), suggesting that the assumption of constant birth and death rates over this time interval is reasonable. Estimates of  $d$  (standard error [SE]) were not significantly different for the two strains: 8.0 (2.14) for CCME 5410 and 7.8 (2.52) for MBIC11017. This corresponds to estimated half-lives (scaled to synonymous site divergence) of 0.087 and 0.089 for *Acaryochloris* strains CCME 5410 and MBIC11017, respectively. That is, most duplicates are expected to be lost rapidly from the genome. These values are within the range observed among eukaryotic genomes (Lynch and Conery 2003).

We estimated the duplicate birth rate  $B$  (the probability that a gene duplicates over the divergence period  $d_s = 0.1$ ) for each genome by  $B = (n_{BD} \times d_s)/N(1 - e^{-d \times d_s})$ , where



**FIG. 3.**—Fraction of duplicates of different divergence levels in the *Acaryochloris* CCME 5410 genome that are shared with the strain MBIC11017 genome.



**FIG. 4.**—Selection ( $d_N/d_S$ ) on duplicates in the strain CCME 5410 (A) and MBIC11017 (B) genomes. Note the different scales on the y axis for (A) and (B).

$n_B$  is the number of duplicate pairs observed at a divergence level below  $d_S = 0.1$ , and  $N$  is the total number of genes in the analysis excluding excess duplicates. The duplicate birth rate of strain MBIC11017 was estimated to be between two and three times greater than that of strain CCME 5410 (0.023 vs. 0.010). We conclude from the above analyses that the observed differences in the frequency distributions of recent duplicate pairs in the genomes of the two strains can be solely explained by differences in their duplication rates.

### Selection on Duplicate Pairs

The idea that redundant gene copies experience a period of relaxed selection (i.e.,  $d_N/d_S \approx 1$ ) following duplication is central to early models of the evolution of novel function (e.g., Ohno 1970). For both *Acaryochloris* genomes, a minority of duplicate pairs does appear to be under relaxed selective constraints immediately following duplication (fig. 4); for duplicates with a divergence level of  $d_S < 0.1$ , mean (SE) values of  $d_N/d_S$  are 0.45 (0.067) and 0.48 (0.084) for the strain MBIC11017 and strain CCME 5410 genomes, with approximately 25% of duplicate pairs having  $d_N/d_S > 0.5$  in both genomes. A small number of these duplicates (four in the strain CCME 5410 genome, nine in the strain MBIC11017 genome) have  $d_N/d_S > 1$ , which suggests that they may be under positive selection. With one exception, the duplication of a chorismate mutase gene in strain MBIC11017, all of these duplicates are annotated as hypothetical or conserved domain proteins. How-

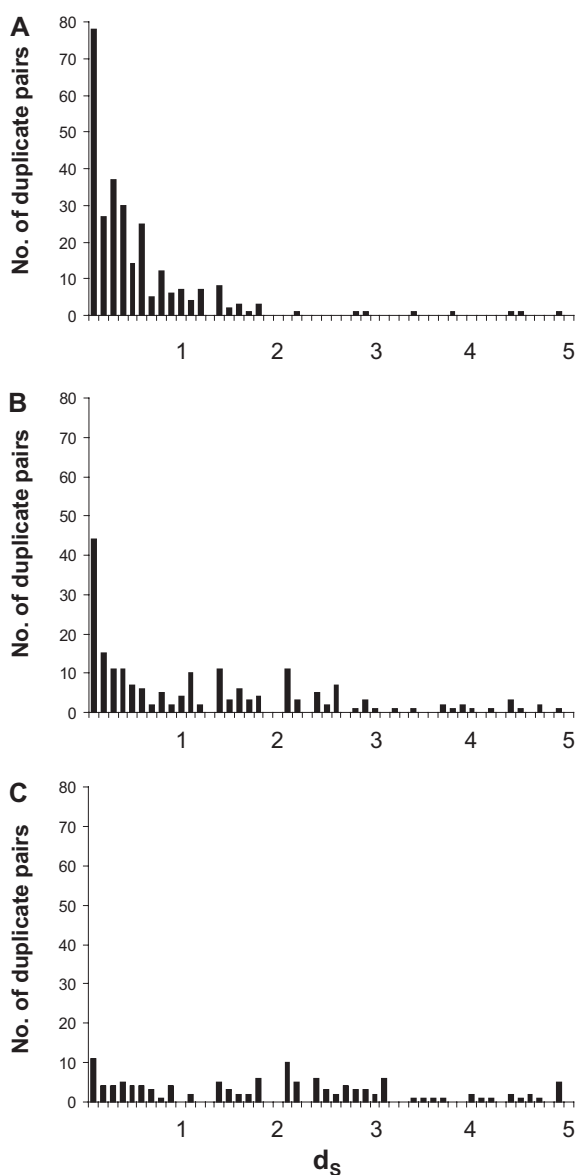
ever, most young duplicates as well as those which have been retained over longer periods appear to be under strong purifying selection against protein change: the median of  $d_N/d_S$  in the  $d_S < 0.1$  divergence level classes is  $\sim 0.2$  and  $\sim 0.3$  for the strain MBIC11017 and strain CCME 5410 genomes, respectively. For duplicates of divergence level greater than  $d_S = 1$ , mean (SE) value of  $d_N/d_S$  is 0.12 (0.004) for the strain MBIC11017 genome and 0.09 (0.004) for the strain CCME 5410 genome. Bearing in mind that the estimated strength of constraint represents the cumulative history of selection since duplication, this pattern indicates that, on average, the intensity of purifying selection on duplicates increases over time. We conclude that the period of near-neutral evolutionary dynamics is at most brief following gene duplication, applies to only a subset of duplicate pairs, and usually is followed by either purging from the genome or an increase in selection against protein change. These results are similar to those obtained for other bacteria (Hooper and Berg 2003b) as well as for eukaryotic genomes (Lynch and Conery 2000, 2003; Aury et al. 2006).

### Physical Location of Duplicated Genes

The location of duplicates at (or near) the time of birth may provide clues regarding the substrates and prevailing mechanisms responsible for duplicate formation. Few duplicates ( $\sim 3\%$  of duplicate pairs) are in tandem (operationally defined here as being within five ORFs of each other) at present in either *Acaryochloris* genome.

The closed genome of *Acaryochloris* strain MBIC 110107 enabled a comprehensive investigation of the distribution of duplicates on the chromosome and on extrachromosomal elements, respectively. For the least divergent classes, at least one gene copy resides on a plasmid for most duplicate pairs (fig. 5A and B), with both on plasmids for greater than 60% of duplicates with a synonymous divergence level of  $d_S < \sim 0.5$ . Because duplicates might move over time, the locations of the least divergent duplicates are expected to be most representative of where they originated. Of the 133 duplicate pairs of divergence level  $d_S < 0.1$ , both members are found on the same genetic element (chromosome or plasmid) only  $\sim 16.5\%$  of the time. Similarly, 13 of 21 identical (i.e.,  $d_S = d_N = 0$ ) duplicate pairs are located on different elements, and of the eight which are on the same element, six likely originated as part of the same duplication event on plasmid pREB3. The origin of most duplicates therefore appears to involve recombination between different plasmids (67/133) or between a plasmid and the chromosome (44/133).

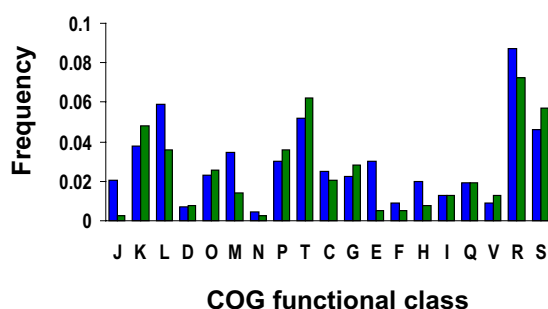
Chromosome–chromosome pairs make a substantial contribution to the pool of duplicates from more divergent classes, however, with 40% of duplicates both residing on the chromosome at divergence levels greater than  $d_S = 2$  (fig. 5C). Plasmid–plasmid duplicates are nearly absent in



**FIG. 5.**—Frequency distributions of duplicate pairs in the *Acaryochloris* strain MBIC11017 genome for duplicate pairs for which both copies currently reside on one or more plasmids (A), one copy is on a plasmid and the other is on the chromosome (B), or both copies are on the chromosome (C).

these classes (fig. 4A). Although most gene duplication events involve interplasmid or plasmid–chromosome exchange, it therefore appears that the vast majority are destined for loss from the genome. Duplicates that are retained over the long term tend to either originate on the chromosome or end up there.

We reach a similar conclusion for the CCMEE 5410 genome (supplementary fig. S1, Supplementary Material online), although we could not assign one or more copies of a duplicate pair to a genetic element for 25% of duplicates. Most of these unassigned pairs belonged to low



**FIG. 6.**—Distributions of clusters of orthologous groups (COG) functional classes for the strain CCMEE 5410 genome (blue) and for duplicate pairs of divergence level  $d_s < 5$  (green).

divergence classes ( $d_s < 1.0$ ), with one-third from a divergence class of  $d_s < 0.1$ . Inability to resolve the locations of these duplicates was due to the presence of one or both copies on a short contig and is likely responsible for the observed lower than expected density of interplasmid and plasmid–chromosome pairs in these low divergence classes (supplementary fig. S1A and B, Supplementary Material online; compare with fig. 2A). The placement on short contigs suggests that they are flanked by repetitive DNA (including IS elements) that may have served as substrates for recombination.

### Duplicate Retention

Bacterial genomes may exhibit a biased retention of duplicates from different gene functional classes (Gevers et al. 2004). Analysis of the strain CCMEE 5410 genome indicated differences in the likelihood of retention among clusters of orthologous groups (COGs) functional classes (fig. 6). In particular, the pool of duplicated genes ( $d_s < 5$ ) is enriched in members from the transcription (K), carbohydrate transport and metabolism (G), ion transport and metabolism (P), signal transduction (T), and unknown (S) functional classes compared with their genome-wide frequencies. Conversely, there is a general paucity of duplicated genes involved in translation (J), replication, recombination and repair (L), cell wall/membrane/envelope biogenesis (M), amino acid transport and metabolism (E), and coenzyme transport and metabolism (H). This suggests that gene dosage balance may generally be more critical within these classes, with duplication of individual genes strongly selected against.

The observed biased retention of recent gene duplications in the G, K, and P classes, as well as a deficiency of H, J, and M classes, is in accord with general longer term evolutionary trends revealed for paralogous gene family expansion in a survey of 48 bacterial genomes (Gevers et al. 2004). The retention of signal transduction (T) and transcription factors (K) is also a feature of plant genomes following polyploidization (Blanc and Wolfe 2004; Maere et al. 2005; Chapman et al. 2006; Thomas et al. 2006).

**Table 2**

Select Strain-Specific Duplicates in the Strain MBIC11017 Genome

ORFs <sup>a</sup>	Annotation	<i>d<sub>s</sub></i>	CCMEE 5410
Nutrient acquisition			
0473/A0147	Fe <sup>2+</sup> -transporter <i>feoB</i>	0.26	7582
0474/A0146	Fe <sup>2+</sup> -transporter <i>feoA</i>	0.11	7581
3038/(B0139/F0079)	Fur transcriptional regulator	0.55/0.30	2939
3040/F0079	Putative Fe <sup>2+</sup> -transporter	0.21	2937
3348/A0161	Fe <sup>3+</sup> -dicitrate ABC transporter	1.36	0699
3349/A0162	Fe <sup>3+</sup> -dicitrate ABC transporter	1.23	0700
3350/A0163	Fe <sup>3+</sup> -dicitrate ABC transporter	1.23	0701
3401/A0182	Fe <sup>3+</sup> -dicitrate ABC transporter	0.22	0727
3402/A0183	Fe <sup>3+</sup> -dicitrate ABC transporter	0.28	0728
3403/A0184	TonB-dependent siderophore transporter	0.15	0729
3416/A0185	Ferrichrome ABC transporter	0.31	0738
C0108/C0205	Heme oxygenase (Fe-recycling)	0	—
3533/3534	Ammonium transporter	0.003	8087
Light harvesting			
1368/3655	Iron deficiency light antenna <i>pcbC</i>	0	8040
C0093/C0216	Phycobilisome linker protein	0.02	—
C0094/C0215	Phycobilisome 32.1 kDa linker	0	—
C0096/C0213	Phycocyanin, alpha subunit	0	—
C0098/C0212	Phycocyanin, beta subunit	0	—
C0099/C0191	Phycocyanin, alpha subunit	0	—
C0100/C0192	Phycocyanin, beta subunit	0.01	—

<sup>a</sup> ORFs on plasmids are preceded by a letter indicating plasmid identity.

Unique duplicates retained by the respective genomes may confer environment-specific fitness benefits through dosage effects, a phenomenon frequently observed in laboratory populations of bacteria (Roth et al. 1996; Romero and Palacios 1997; Reams and Neidle 2003). A correlation between duplicate content and environment has also been observed for yeast (Ames et al. 2010). The genome of strain MBIC11017 possesses a striking number of duplicated genes involved in nutrient acquisition (principally the binding, transport, and metabolism of iron) that exist as either single copies or are not found in the strain CCMEE 5410 genome (table 2). All but one of these include a plasmid-encoded duplicate copy. We note that the strain MBIC11017 genome also includes eight plasmid-encoded single-copy iron acquisition genes that are absent from the strain CCMEE 5410 genome (ORFs A0156, A0157, A0172, A0197, A0198, A0274, B0123, B0125).

That this strain's genome may have been shaped by iron limitation is also suggested by the recent duplication of the light antenna protein *pcbC* (table 2). This gene is upregulated by *Acaryochloris* cells under conditions of iron deficiency (Chen et al. 2005), and PcbC protein subunits produce a light-harvesting antenna for photosystem I that may compensate for the reduction in the level of this photosystem relative to photosystem II that occurs during iron stress.

Tropical Pacific waters generally appear to be low in iron (e.g., Coale et al. 1996; Behrenfeld et al. 2006). Although we do not know the iron concentration of the local environment from which strain MBIC11017 was isolated, there are

reasons to believe that *Acaryochloris* cells may be iron limited in their natural habitat. This strain was isolated from underneath the ascidian, *Lissoclinum patella* (Miyashita et al. 1996), which belongs to a suborder (Aplousobranchia), which includes members notable for the accumulation of high concentrations of iron from the environment in blood cells called ferrocytes (Endean 1955). In addition, the positive response of MBIC11017 laboratory cultures to heavy iron addition suggests an organism with high demand for this nutrient (Swingley et al. 2005).

Other recent duplicates in the MBIC11017 genome that are involved in light-harvesting encode pigment and scaffold components of phycobiliproteins (table 2), the major accessory pigments in photosynthesis for most cyanobacteria. Multiple duplicate copies of genes for the phycobiliprotein phycocyanin, which specifically harvests yellow–orange light for photosynthesis, as well as linker proteins essential for the assembly of phycobiliprotein rods, are located on plasmid pREB3 (Swingley et al. 2008). Strain MBIC11017 produces phycobiliproteins under low light conditions in the laboratory (Chan et al. 2007). By contrast, strain CCMEE 5410 does not produce phycobiliproteins (Chan et al. 2007), and these genes are missing entirely from its genome (table 2). This pattern suggests differences in the availability of yellow–orange light in the two environments. These wavelengths appear to be available at low levels in the natural environment of strain MBIC11017 (Kühl et al. 2005), whereas they may be more rapidly attenuated in the turbid Salton Sea environment from which strain CCMEE 5410 was

**Table 3**

Select Strain-Specific Duplicates in the Strain CCME 5410 Genome

ORFs <sup>a</sup>	Annotation	<i>d<sub>s</sub></i>	MBIC11017
Carbon metabolism			
0720/(2355/2491)	Fructose-bisphosphate aldolase	2.59/1.48	3372
2488/(1772/2358)	Xu5P/Fru6P phosphoketolase	2.69/2.27	0443
1774/2356	Acetate kinase	2.57	0445
2357/2490	Phosphoglycerate mutase family	1.58	—
4274/5615	Phosphoglycerate mutase family	1.03	—
2365/2496	Putative glycogen phosphorylase	0.84	—
Copper resistance			
2343/2458	Cu resistance protein CopA	0.31	—
2364/2487	Copper-translocating ATPase	1.10	—
2372/2481	Copper-translocating ATPase	1.21	—
Defense mechanisms			
3189/7004	RND family multidrug efflux	2.22	2480
1784/6258	RND family multidrug efflux	0.48	0454
Redox homeostasis			
2586/(2383/2469)	Glutaredoxin	1.69/0.84	3463

<sup>a</sup> ORFs assigned to plasmids are italicized.

isolated (Miller et al. 2005) by phycobiliprotein-producing plankton (Wood et al. 2002) and inorganic particulate matter (Swan et al. 2007) in the overlaying water column.

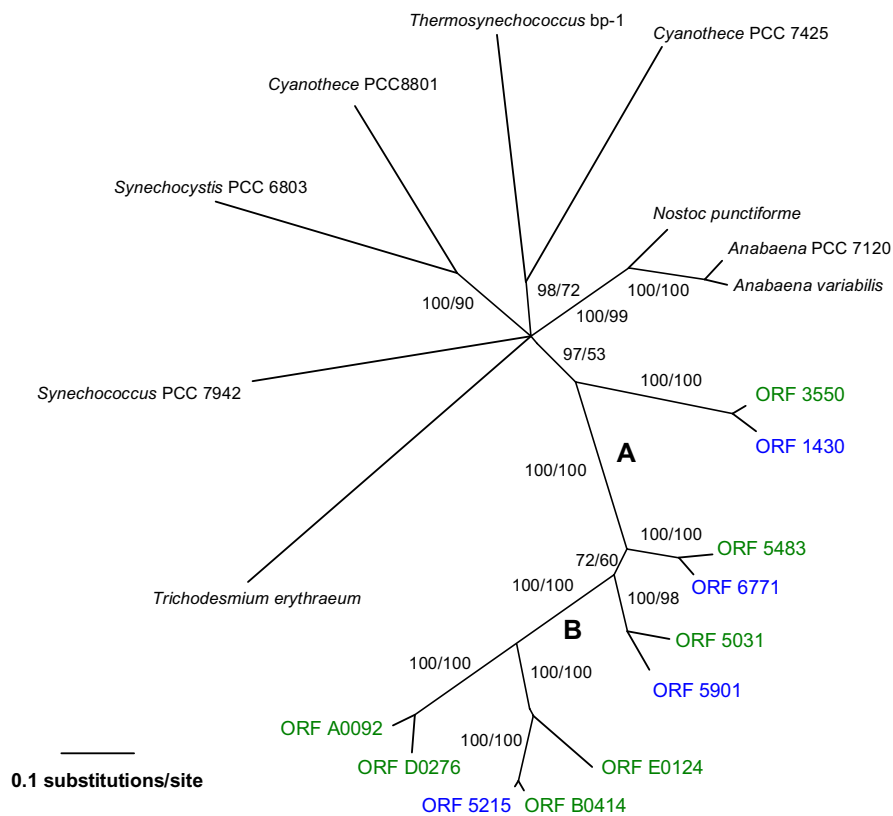
The Salton Sea is a phosphorus-limited system characterized by high concentrations of dissolved organic carbon and nitrogen (Schroeder et al. 2002), as well as of iron, primarily as reduced particulates (Holdren and Montañó 2002; de Koff et al. 2008). Heavy metal (including cadmium, copper, selenium, and zinc) concentrations are also high (Vogl and Henry 2002; LeBlanc and Schroeder 2008). The unique duplicate pool of the strain CCME 5410 genome (table 3) is enriched in loci involved in organic carbon metabolism and heavy metal resistance (in particular, copper). Many duplicate copies were found on two inferred plasmid contigs encoding ORFs 2340–2395 and ORFs 2456–2512, respectively. Similarly, a number of single-copy genes in the strain CCME 5410 genome in large (>50 kbp), plasmid-assigned regions of no apparent homology with strain MBIC11017 are also involved in heavy metal (primarily copper) resistance (ORFs 2382 [copper-translocating ATPase], 2458 [CopA family copper-resistance protein], 7833 [copper-resistance protein precursor CopB], 0016 [CzcA family heavy metal efflux pump]).

### A Role for *recA* Dosage in Gene Duplication?

A mechanistic understanding of the gene duplication dynamics of *Acaryochloris* genomes must ultimately account for both their high load of recent duplicates compared with other bacteria and the observed differences in the duplicate age distributions of the two strains. The recombination process is a likely candidate for involvement in duplication because homologous recombination functions are generally important for both duplicate formation (Hill et al. 1977; Dimpfl and Echols 1989; Petit et al. 1991) and loss (Anderson

and Roth 1979), particularly for long recombining sequences such as IS elements (but see Reams et al. [2010], e.g., in which duplication depends only weakly on homologous recombination). The large number of IS elements in *Acaryochloris* genomes (table 1) provide potential substrates for recombination. Although there appears to be a general trend against the retention of duplicates involved in DNA replication, recombination, and repair (fig. 6), both genomes contain a number of duplicated genes from this functional class, and we briefly consider here whether these duplicates may play a role in the enhanced duplication dynamics of these genomes.

Most notably, there are an unusually large number of *recA* copies in both *Acaryochloris* genomes. RecA is a multifunctional protein that is central to homologous recombination, is involved in recombination-mediated DNA damage repair and rescue of stalled replication forks, is required for mutagenesis mediated by translesion synthesis, and regulates gene expression through its coprotease activity (reviewed by Miller and Kokjohn 1990). The strain MBIC11017 genome contains seven *recA* copies (Swingley et al. 2008), whereas there are four complete copies in the genome of strain CCME 5410. The CCME 5410 genome also includes a truncated copy (ORF 6290) with a nonsense mutation at codon 241 produced by an apparent transposition event that results in the loss of part of the ATPase core and the C-terminal domain; the putative 3' end of the gene copy is found on a different contig (ORF 8203) and is also adjacent to a transposase. In contrast, *recA* exists as a single copy in the vast majority of bacterial genomes; the only known exceptions are the *Acaryochloris* genomes and those of *Myxococcus xanthus* (two copies; Norioka et al. 1995), *Bacillus megaterium* (two copies; Nahrstedt et al. 2005), and *Deinococcus deserti* (three copies; de Groot et al. 2009).



**Fig. 7.**—Unrooted Bayesian phylogeny of *Acaryochloris recA* duplicates. Values at a node represent the Bayesian clade credibility followed by the bootstrap value for a ML analysis. MBIC111017 copies are green and CCMEE 5410 copies are blue.

*Escherichia coli* exhibits a 10-fold or greater tandem duplication rate if RecA is constitutively activated (Dimpfl and Echols 1989), and overexpression of its eukaryotic homolog RAD51 may also enhance duplication rate as well as generally increase genome instability (reviewed by Klein 2008). Whether the greater *recA* copy number in *Acaryochloris* genomes results in enhanced expression remains to be determined, but the association between copy number and strain duplication rate is consistent with a dosage effect. Also consistent with this possibility, the *D. deserti* genome likewise appears to contain a greater number of paralogs (~100–200) than those of its single-copy congeners, *D. radiodurans* and *D. geothermalis* (de Groot et al. 2009).

*Acaryochloris recA*s are both extremely diverse and monophyletic, indicating that this diversity likely originated solely during *Acaryochloris* diversification rather than by horizontal gene transfer (HGT) (fig. 7). Three chromosomal copies are shared by the strains and appear to predate divergence from their common ancestor, whereas the strains vary in the number of plasmid-borne copies. Although on average all copies have experienced strong purifying selection ( $d_N/d_S = 0.05$ ), there is some evidence that certain amino acid substitutions have been selectively favored during *recA* diversification. Along two branches (labeled A and B in fig. 7), branch-site models of codon evolution (Yang and Nielsen 2002) which

allow for positive selection on one or a few codon sites on specific branches of a phylogeny had significantly greater likelihood values than nearly-neutral models constrained to  $d_N/d_S \leq 1$  for all codons ( $2\Delta L = 70.42$ ,  $P = 0$  for the Branch A model;  $2\Delta L = 9.16$ ,  $P = 0.01$  for the Branch B model). The codons estimated to have experienced positive selection (i.e.,  $d_N/d_S > 1$  with a posterior probability  $P > 0.95$  by Bayesian analysis) at some point during *recA* diversification (supplementary fig. S2; Supplementary Material online) include sites that participate in monomer–monomer interactions in the RecA filament (codons 105, 114, 115, 127, 153, and 240), make contact with ssDNA-binding sites (codon 153), or change the properties (e.g., charge) of the C-terminal domain of the protein (codons 312, 323, and 328), which is known to autoregulate RecA activity and to bind dsDNA during homologous recombination (Cox 2007). Whether these changes have consequences for RecA structure and function remains to be investigated, as does the possibility that diversification has yielded paralogous RecAs with nonredundant functions (i.e., subfunctionalization) in *Acaryochloris* cells.

## Concluding Remarks

Strain-specific duplicates concentrated on plasmids make a substantial contribution to gene content differences

between *Acaryochloris* genomes and appear to be selectively retained in their respective contemporary environments by favorable dosage effects. These differences are in part the product of the differential retention of duplicates of chromosomal origin (fig. 4B; see below). The lower degree of conservation of gene content on plasmids compared with the chromosome also suggests an important role for HGT in *Acaryochloris* evolution. If this is the case, the implication is that the ultimate source of many duplicate pairs is a single-copy gene of foreign origin. In Proteobacteria and Firmicutes, horizontally transferred genes do appear to be more likely to be duplicated (Hooper and Berg 2003a).

To obtain a conservative estimate of the contribution of HGT to the pool of strain-specific duplicates with at least one copy on a plasmid, we performed BlastP analyses against the NCBI Blast nonredundant protein sequence database for each genome. For a given strain-specific duplicate family without an ortholog in the other strain, it can be difficult to unequivocally determine whether it is the product of the differential retention of an ancestral gene or of HGT. This is because many of these loci either exhibit greatest sequence similarity to a different cyanobacterium or have no similarity to another sequence in the database (i.e., are orphan genes). Therefore, taking a conservative approach and using an *E* cutoff value of  $10^{-20}$ , we considered a duplicate family to be of vertical origin if the top non-*Acaryochloris* hit for a duplicate family was a cyanobacterium, to have originated by HGT if the top hit was another taxon and to be of unknown origin if it was an orphan.

For duplicate pairs for which one copy is on the chromosome (fig. 4B), most are inferred to be of cyanobacterial origin in both *Acaryochloris* genomes by the above criteria (62% for strain MBIC11017 and 77% for the subset of duplicates in strain CCME 5410 which could be fully assigned to genetic elements). This is the expectation if the plasmid copy was derived by duplication of a chromosomal template. Fewer duplicates in this category appear to involve horizontally transferred loci (4% and 3%, respectively). For interplasmid duplicates, however, a larger fraction shows highest similarity to a taxon other than a cyanobacterium and likely owes its origins to HGT (8% and 14%, respectively). For example, CCME 5410-specific duplicate pairs ORF2364/ORF2487 and ORF2365/ORF2496 (table 3) exhibit greatest sequence identities to *Thermus thermophilus* (67%) and a planctomycete bacterium (60%), respectively, and the former has no known homolog among other cyanobacteria. We believe that these HGT estimates are probably very low, as approximately half of the duplicates in the interplasmid category were orphans (57% and 50%, respectively) which may be the products of HGT. We conclude that the atypical, largely plasmid-mediated duplication dynamics of *Acaryochloris* genomes generate copy number variation among loci of both ancestral and

foreign origin, that this variation is frequently nonadaptive, but that it also is an important source of locally adaptive genomic variation with the potential to rapidly respond to environmental change.

In addition to modifying gene dosage, duplication also creates opportunities for the evolution of novel gene functions. Whether neofunctionalization contributed to the innovation of the unique chlorophyll metabolism of *Acaryochloris* remains unresolved, as the details of Chl *d* biosynthesis and degradation are yet to be fully elucidated. Chl *d* differs from Chl *a* by the replacement of a vinyl group with a formyl group at C-3 of the porphyrin ring. The pigment is produced from Chl *a* and molecular oxygen precursors (Schliep et al. 2010), and biochemical evidence suggests that the “Chl *d* synthase” that performs this reaction is a P450 oxygenase (Chen 2010). The genomes of both strains each harbor ten genes encoding P450 enzymes; however, none of the copies appear to be recent duplicates (not shown). We analyzed the pool of duplicates retained by both genomes for paralogs with homology to other proteins that could potentially participate in other aspects of Chl *d* metabolism such as porphyrin ring degradation. The only candidates to emerge were a pair of divergent ( $d_s \approx 1.9$ ) duplicates with homology to a family of Rieske-FeS motif-containing oxygenases involved in chlorophyll synthesis and degradation (ORFs 0307/5640 in CCME 5410 and 0159/A0067 in MBIC11017). It is notable that A0067 is found within one of the few regions of extensive synteny between a MBIC11017 plasmid and the CCME 5410 genome (including A0036–A0053 and A0066–A0075). Whether one of these paralogs has diverged to specifically degrade Chl *d* awaits further investigation.

## Supplementary Material

Supplementary table S1 is available at *Genome Biology and Evolution* online (<http://www.gbe.oxfordjournals.org/>).

## Acknowledgments

We gratefully acknowledge the support of the Gordon and Betty Moore Foundation's Marine Microbiology Initiative. We also thank Robert Montgomery and Justin Johnson for their contributions to the CCME 5410 genome project annotation and coordination, respectively. The paper benefited from the comments of John McCutcheon, Michael Weltzer, and two anonymous reviewers. This research was supported by National Science Foundation award EF-0801999 to S.R.M.

## Literature Cited

- Altschul SF, Gish W, Miller W, Myers E, Lipman D. 1990. Basic local alignment search tool. *J Mol Biol.* 215:403–410.
- Ames RM, et al. 2010. Gene duplication and environmental adaptation within yeast populations. *Genome Biol Evol.* 2:591–601.

- Anderson RP, Roth JR. 1977. Tandem genetic duplications in phage and bacteria. *Annu Rev Microbiol.* 31:473–505.
- Anderson RP, Roth JR. 1979. Gene duplication in bacteria: alteration of gene dosage by sister-chromosome exchanges. *Cold Spring Harb Symp Quant Biol.* 43:1083–1087.
- Aury J-M, et al. 2006. Global trends of whole-genome duplications revealed by the ciliate *Paramecium tetraurelia*. *Nature.* 444:171–178.
- Behrendt L, et al. 2011. Endolithic chlorophyll *d*-containing phototrophs. *ISME J.* 5:1072–1076.
- Behrenfeld MJ, et al. 2006. Controls on tropical Pacific Ocean productivity revealed through nutrient stress diagnostics. *Nature.* 442:1025–1028.
- Blanc G, Wolfe KH. 2004. Functional divergence of duplicated genes formed by polyploidy during *Arabidopsis* evolution. *Plant Cell.* 16:1679–1691.
- Chan Y-W, et al. 2007. Pigment composition and adaptation in free-living and symbiotic strains of *Acaryochloris marina*. *FEMS Microbiol Ecol.* 61:65–73.
- Chapman BA, Bowers JE, Feltus FA, Paterson AH. 2006. Buffering of crucial functions by paleologous duplicated genes may contribute cyclicity to angiosperm genome duplication. *Proc Natl Acad Sci U. S. A.* 103:2730–2735.
- Chen M. 2010. Spectral extension of photosynthesis using Chlorophyll *d*. In *Vitro Cell Dev Biol Anim.* 46:S14–S15.
- Chen M, Bibby TS, Nield J, Larkum A, Barber J. 2005. Iron deficiency induces a chlorophyll *d*-binding Pcb antenna system around Photosystem I in *Acaryochloris marina*. *Biochim Biophys Acta.* 1708:367–374.
- Coale KH, et al. 1996. A massive phytoplankton bloom induced by an ecosystem-scale iron fertilization experiment in the equatorial Pacific Ocean. *Nature.* 383:495–501.
- Coissac E, Maillier E, Netter P. 1997. A comparative study of duplications in bacteria and eukaryotes: the importance of telomeres. *Mol Biol Evol.* 14:1062–1074.
- Cox MM. 2007. Regulation of bacterial RecA protein function. *Crit Rev Biochem Mol Biol.* 42:41–63.
- de Groot A, et al. 2009. Alliance of proteomics and genomics to unravel the specificities of Sahara bacterium *Deinococcus deserti*. *PLoS Genet.* 5:e1000434.
- de Koff J, Anderson M, Amrhein C. 2008. Geochemistry of iron in the Salton Sea, California. *Hydrobiologia.* 604:111–121.
- Delcher AL, Harmon D, Kasif S, White O, Salzberg SL. 1999. Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* 27:4636–4641.
- Dimpfl J, Echols H. 1989. Duplication mutation as an SOS response in *Escherichia coli*: enhanced duplication formation by a constitutively activated RecA. *Genetics.* 123:255–260.
- Endean R. 1955. Studies of the blood and tests of some Australian ascidians. I. The blood of *Pyura stolonifera* (Heller). *Aust J Mar Freshwat Res.* 6:157–164.
- Finn RD, et al. 2008. The Pfam protein families database. *Nucleic Acids Res.* 36:D281–D288.
- Gevers D, Vandepoele K, Simillion C, Van de Peer Y. 2004. Gene duplication and biased functional retention of paralogs in bacterial genomes. *Trends Microbiol.* 12:148–154.
- Haack KR, Roth JR. 1995. Recombination between chromosomal IS200 elements supports frequent duplication formation in *Salmonella typhimurium*. *Genetics.* 141:1245–1252.
- Haft DH, Selengut JD, White O. 2003. The TIGRFAMs database of protein families. *Nucleic Acids Res.* 31:371–373.
- Hill CW, Grafstrom RH, Harnish BW, Hillman BS. 1977. Tandem duplications resulting from recombination between ribosomal RNA genes in *Escherichia coli*. *J Mol Biol.* 116:407–428.
- Holdren GC, Montañó A. 2002. Chemical and physical characteristics of the Salton Sea, California. *Hydrobiologia.* 473:1–21.
- Hooper SD, Berg OG. 2003a. Duplication is more common among laterally transferred genes than among indigenous genes. *Genome Biol.* 4:R48.
- Hooper SD, Berg OG. 2003b. On the nature of gene innovation: duplication patterns in microbial genomes. *Mol Biol Evol.* 20:945–954.
- Huelsenbeck JP, Ronquist F. 2001. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics.* 17:754–755.
- Irish VF, Litt A. 2005. Flower development and evolution: gene duplication, diversification and redeployment. *Curr Opin Genet Dev.* 15:454–460.
- Jordan I, Makarova K, Spouge J, Wolf Y, Koonin E. 2001. Lineage-specific gene expansions in bacterial and archaeal genomes. *Genome Res.* 11:555–565.
- Kashiyama Y, et al. 2008. Evidence of global chlorophyll *d*. *Science.* 321:658.
- Klein HL. 2008. The consequences of Rad51 overexpression for normal and tumor cells. *DNA Repair.* 7:686–693.
- Kühl M, Chen M, Ralph PJ, Schreiber U, Larkum AWD. 2005. Ecology: a niche for cyanobacteria containing chlorophyll *d*. *Nature.* 433:820.
- Kurtz S, et al. 2004. Versatile and open software for comparing large genomes. *Genome Biol.* 5:R12.
- LeBlanc L, Schroeder R. 2008. Transport and distribution of trace elements and other selected inorganic constituents by suspended particulates in the Salton Sea Basin, California, 2001. *Hydrobiologia.* 604:123–135.
- Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25:955–964.
- Lynch M, Conery JS. 2000. The evolutionary fate and consequences of duplicate genes. *Science.* 290:1151–1155.
- Lynch M, Conery JS. 2003. The evolutionary demography of duplicate genes. *J Struct Funct Genomics.* 3:35–44.
- Maere S, et al. 2005. Modeling gene and genome duplications in eukaryotes. *Proc Natl Acad Sci U. S. A.* 102:5454–5459.
- Makarova KS, Wolf YI, Mekhedov SL, Mirkin BG, Koonin EV. 2005. Ancestral paralogs and pseudoparalogs and their role in the emergence of the eukaryotic cell. *Nucleic Acids Res.* 33:4626–4638.
- Miller RV, Kokjohn TA. 1990. General microbiology of *recA*: environmental and evolutionary significance. *Annu Rev Microbiol.* 44:365–394.
- Miller SR, et al. 2005. Discovery of a free-living chlorophyll *d*-producing cyanobacterium with a hybrid proteobacterial/cyanobacterial small-subunit rRNA gene. *Proc Natl Acad Sci U. S. A.* 102:850–855.
- Miyashita H, et al. 1996. Chlorophyll *d* as a major pigment. *Nature.* 383:402–402.
- Mohr R, et al. 2010. A new chlorophyll *d*-containing cyanobacterium: evidence for niche adaptation in the genus *Acaryochloris*. *ISME J.* 4:1456–1469.
- Nahrstedt H, Schroder C, Meinhardt F. 2005. Evidence for two *recA* genes mediating DNA repair in *Bacillus megaterium*. *Microbiology.* 151:775–787.
- Norioka N, Hsu MY, Inouye S, Inouye M. 1995. Two *recA* genes in *Myxococcus xanthus*. *J Bacteriol.* 177:4179–4182.
- Ohno S. 1970. Evolution by gene duplication. New York: Springer-Verlag.

- Petit MA, Dimpfl J, Radman M, Echols H. 1991. Control of large chromosomal duplications in *Escherichia coli* by the mismatch repair system. *Genetics*. 129:327–332.
- Posada D, Crandall KA. 1998. MODELTEST: testing the model of DNA substitution. *Bioinformatics*. 14:817–818.
- Reams AB, Kofoed E, Savageau M, Roth JR. 2010. Duplication frequency in a population of *Salmonella enterica* rapidly approaches steady state with or without recombination. *Genetics*. 184:1077–1094.
- Reams AB, Neidle EL. 2003. Genome plasticity in *Acinetobacter*: new degradative capabilities acquired by the spontaneous amplification of large chromosomal segments. *Mol Microbiol*. 47:1291–1304.
- Romero D, Palacios R. 1997. Gene amplification and genomic plasticity in prokaryotes. *Annu Rev Genet*. 31:91–111.
- Roth J, et al. 1996. Rearrangements of the bacterial chromosome: formation and applications. In: Neidhardt F, Curtis III R, Ingraham JL, Lin ECC, Low KB, Magasanik B, Reznikoff WS, Riley M, Schaechter M, Umberger HE, editors. *Escherichia coli* and *Salmonella typhimurium*: cellular and molecular biology. Washington (DC): ASM Press.
- Schliep M, Crossett B, Willows RD, Chen M. 2010. 18O labeling of Chlorophyll *d* in *Acaryochloris marina* reveals that Chlorophyll *a* and molecular oxygen are precursors. *J Biol Chem*. 285:28450–28456.
- Schroeder RA, Orem WH, Kharaka YK. 2002. Chemical evolution of the Salton Sea, California: nutrient and selenium dynamics. *Hydrobiologia*. 473:23–45.
- Swan BK, et al. 2007. Spatial and temporal patterns of transparency and light attenuation in the Salton Sea, California, 1997–1999. *Lake Reserv Manage*. 23:653–662.
- Swingley WD, Hohmann-Marriott MF, Olson TL, Blankenship RE. 2005. Effect of iron on growth and ultrastructure of *Acaryochloris marina*. *Appl Environ Microbiol*. 71:8606–8610.
- Swingley WD, et al. 2008. Niche adaptation and genome expansion in the chlorophyll *d*-producing cyanobacterium *Acaryochloris marina*. *Proc Natl Acad Sci U. S. A*. 105:2005–2010.
- Swofford DL. 1996. PAUP 3.1.1. Sunderland (MA): Sinauer Associates.
- Taylor J, Raes J. 2004. Duplication and divergence: the evolution of new genes and old ideas. *Annu Rev Genet*. 38:615–643.
- Thomas BC, Pedersen B, Freeling M. 2006. Following tetraploidy in an *Arabidopsis* ancestor, genes were removed preferentially from one homeolog leaving clusters enriched in dose-sensitive genes. *Genome Res*. 16:934–946.
- Thompson J, Higgins D, Gibson T. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. *Nucleic Acids Res*. 22:4673–4680.
- True JR, Carroll SB. 2002. Gene co-option in physiological and morphological evolution. *Annu Rev Cell Dev Biol*. 18:53–80.
- Vogl RA, Henry RN. 2002. Characteristics and contaminants of the Salton Sea sediments. *Hydrobiologia*. 473:47–54.
- Wagner A. 2008. Gene duplications, robustness and evolutionary innovations. *Bioessays*. 30:367–373.
- Wood AM, Miller SR, Li W, Castenholz RW. 2002. Preliminary studies of cyanobacteria, picoplankton, and virioplankton in the Salton Sea with special attention to phylogenetic diversity among eight strains of filamentous cyanobacteria. *Hydrobiologia*. 473:77–92.
- Yang Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci*. 13:555–556.
- Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol Biol Evol*. 19:908–917.

**Associate editor:** Martin Embley